Cost Functions From Cross-Section Data—Fact or Fantasy?

By J. F. Stollsteimer, R. G. Bressler and J. N. Boles

Production and cost functions have long been recognized as vital components of economic analyses relating to the individual firm. The U.S. Department of Agriculture, beginning with the pioneering work of W. J. Spillman, has been a continuing participant in their empiric and theoretical development. Whereas early work emphasized farm production and cost functions, much attention has centered lately on the marketing firm. This attention has brought into sharper focus certain organizational and operating characteristics of plants. With growing interest in the marketing area, the work in the Department expanded to include cooperative research with several State experiment stations. A major such effort has involved the Marketing Economics Division, Economic Research Service, and the California Agricultural Experiment Station. This is the first of three papers prepared for publication in Agricultural Economics Research to reflect some aspects of theoretical and methodological developments in these studies. The following paper comments on, and extends the results of, a statistical analysis of costs in the operation of feed mills developed in a cooperative study with the Iowa Agricultural Experiment Station, and reported in this journal by Richard Phillips in 1956. In a second paper the authors will deal with the possibilities of developing production and cost functions from more detailed analysis of accounting records of individual firms. A third paper will discuss the technique of plant cost synthesis. *This report grew out of research in plant cost and efficiency carried on cooperatively by the Marketing Economics Division, Economic Research Service, and the Giannini Foundation of Agricultural Economics, University of California at Berkeley. The authors are indebted to L. L. Sammet, B. C. French, and D. B. DeLoach of the University of California, and W. F. Finner and V. J. Brensike of the Economic Research Service, U.S. Department of Agriculture, for their helpful suggestions during the preparation of this paper.*

WITH THE CURRENT emphasis on cost and efficiency research, attempts to develop empirical representations of short- and long-run cost functions are both common and important. Two principal approaches to quantification are: (1) the synthetic method—building up descriptions of cost functions from detailed study of plant stages and operations and the integration of these stages to represent the total plant operation, and (2) the statistical approach, deriving relationships from the analysis of aggregate cost and volume data.

The synthetic method frequently involves relatively large and expensive research inputs. It is suspect in some quarters because of the "unreal" connotation of its name. The statistical approach, on the other hand, frequently utilizes readily available "cross-section" data. It can, therefore, produce results with relatively small research cost, and it has the added appeal of reflecting "real" plant operations. Furthermore, the regression coefficients obtained can be subjected to statistical tests of reliability, though this may be an advantage of dubious value.

This paper reports on a series of pragmatic explorations of the nature of results obtained by the statistical analysis of cross-section data. It has its immediate origin in a report by Richard Phillips of Iowa State University published in 1956, and, in a real sense, is a continuation of the methodological inquiry of that paper.[1] The authors are indebted to Professor Phillips for his assistance in making available details of his data and of revised analyses, as well as for his critical comments on earlier versions of this report.

[1] Richard Phillips, "Empirical Estimates of Cost Functions for Mixed Feed Mills in the Mid-West," *Agricultural Economics Research*, vol. VIII, no. 1, January, 1956, pp. 1–8.

Reprinted by the Economic Research Service from AGRICULTURAL ECONOMICS RESEARCH, Vol. XIII, No. 3, July 1961.

ERS-27

September 1961

The General Approach

Basic data for the investigation covered in this report are from the Phillips study of 29 feed mills. They include (1) V—total annual volume in tons of feed mixed; (2) C—total annual costs for each plant; and (3) K—annual capacity in tons of feed mixed.[2] Annual volume and capacity data can be combined in appropriate ways to define other related variables such as excess capacity, capacity rates per hour, or equivalent full-time hours or days of operation per year. These data are more or less typical of cross-section data available for samples of marketing or processing plants, although the capacity information represents a somewhat unusual and strategic addition.

These data were used in all formulations reported in this paper. The basic approach was to select a number of alternative models or type equations for the cost relationship, apply these to the single set of data, and, finally, to compare and contrast the results. All models use both volume and capacity, directly or indirectly, as independent variables in the regression analyses and total annual costs as the dependent variable.

For any of the models used, the application of multiple regression techniques results in an equation relating total annual cost to annual volume and capacity. Both short- and long-run cost functions may then be obtained from the multiple regression equation.

Short-run functions are described by specifying alternative levels of capacity—each assumed to be associated with a specific but undefined fixed plant—and then relating total annual cost to annual volume for volume up to but not exceeding the selected annual capacity.

A long-run cost function is computed for each model by specifying alternative annual volumes and then selecting for each volume the plant capacity which will minimize costs, subject to the condition that annual capacity is greater than, or equal to, the selected annual volume.

[2] Estimates of capacity were based on actual peak weekly output during past plant operations, rather than on engineering measurements. Peak weekly volume was divided by the corresponding weekly hours of plant operation to obtain an estimate of capacity output rate per hour. Annual capacity was defined in terms of operations for 22.5 hours per day, 6 days per week, and 52 weeks per year. Actual plant volumes ranged from 466 to 141,775 tons per year. Plant capacities for the sample plants ranged from 7,020 to 585,000 tons per year.

In some cases, the resulting long-run equation can be obtained directly from the regression equation by setting capacity equal to volume or excess capacity equal to zero—these correspond to the familiar "J-shaped" average cost curves found in empirical analyses. In others, costs are minimized by selecting capacities somewhat in excess of the selected volumes—these represent the tangency solutions emphasized in Viner's classic paper.

Four general types of models were used in these investigations, with a number of specific forms: (1) the original Phillips model of general form $C=b_1V^n+b_2(K-V)$: (2) a modification of the Phillips model with nonzero intercept in the general form $C=a+b_1V^n+b_2(K-V)$; (3) a series of models involving constant marginal costs for any short-run function and with fixed costs increasing with capacity, representing various elaborations of the form $C=(a+b_1K)+(b_2+b_3K)V$; and (4) an illustration of a form developed graphically as an envelope function rather than by conventional regression techniques. The specific forms used are discussed in the paragraphs that follow.

Model 1.—This is the original Phillips form, and Equations 1a, 1b, 1c, and 1d—(table 1)—are the Phillips results obtained for specific values for n of 0.5, 0.7, 0.8, and 0.9. Phillips apparently chose this form on the basis of two of its properties: (1) it yields a total cost function which increases at a decreasing rate and (2) passes through the origin. He states that "such a model is logical because total costs should be zero when both output and unused capacity are zero."[3] In addition, Equations 1e and 1f have been fitted, using n values of 1.0 and 1.1. For all of these equations, the long-run cost function is obtained by specifying particular values for V and determining the values for capacity K which will minimize costs, subject to the conditions that $V>0$, $K>0$, and $K\geqq V$. The change in total cost with respect to change in capacity is given by the partial derivative, or:

$$\frac{\partial C}{\partial K}=b_2$$

If b_2 is positive—as it must be to be logically admissible—the total cost of producing any volume V will be minimized by making capacity K as

[3] Phillips, op. cit., p. 5.

small as possible, that is $K=V$. The long-run cost function thus reduces to:

$$C=b_1V^n, \qquad V=K,^4$$

A function that passes through the origin and which will show economies of scale when n has values less than 1.0, constant returns to scale when n equals 1.0, and diseconomies of scale when n has values greater than 1.0.

Short-run or plant cost functions are obtained from the multiple relationship by assigning any constant value for capacity, say $\bar{K}$, and expressing C as a function of V:

$$C=b_1V^n+b_2(\bar{K})-V$$

Short-run marginal costs are then defined, with K fixed at $\bar{K}$, by the derivative:

$$\frac{dC}{dV}=nb_1V^n-1-b_2, \qquad V\leqq\bar{K}$$

When n is positive but less than 1.0, short-run marginal costs decline monotonically with increases in volume and eventually become negative. This must be regarded as questionable on a priori grounds even in the ranges where marginal costs are positive and, of course, is quite unacceptable in volume ranges where the indicated marginal costs are negative. That is to say, we must reject on logical grounds a total cost curve for a plant which increases with volume to a maximum and then decreases—suggesting that the plant could produce larger volumes for lower total cost than some smaller volumes. When n equals 1.0, indicated marginal short-run costs are constant regardless of volume—the total cost curve is linear. When n is greater than 1.0 but less than 2.0, marginal costs increase throughout the entire range of volume, but at a decreasing rate—a peculiar form when compared to usual theoretical constructs.

Model 2.—This formulation is identical with the original Phillips model except that the functions are not forced through the origin. This modification was made after it was observed that the original equations generally overestimated costs for the smaller plants. Equations 2a, 2b, 2c, and 2d have assigned n values of 0.5, 0.7, 0.8, and 0.9 and so correspond to original Equations 1a through 1d. Notice that, if the fitted intercept values are negative, the indicated total costs will be negative for very small values of V and K. While this is obviously unacceptable, the form may give good descriptive "fits" over most of the relevant ranges.

Model 3.—A priori reasoning about the operation of mechanized processing plants suggests that short-run marginal costs will be constant if equipment operates at fixed output rates and if annual volume is varied, either by varying the number of parallel lines operated, or by varying the number of hours of plant operation. It is also reasonable to expect that fixed costs will be an increasing function of capacity. If plant capacity is increased by the addition of identical, parallel lines, fixed costs should increase in an approximately linear (but discontinuous) relation with capacity, while short-run marginal costs should be unaffected by plant size. On the other hand, if capacity is increased by using larger and larger items of equipment, short-run marginal costs would probably decrease with increases in capacity. The specific equations selected to reflect these possibilities are:

3a: $C=a+b_1V+b_2K$

3b: $C=(a+b_1K)+(b_2+b_3K)V$

3c: $C=(b_1K)+(b_2+b_3K)V$

3d: $C=(a+b_1K+b_2K^2)+(b_3+(b_4K+b_5K^2)V$

Notice that, for all of these equations, short-run marginal costs are constant. That is, for any fixed capacity, short-run marginal costs are not a function of volume. For Equation 3a, short-

[4] This function has the following properties depending upon the value of n:

	$0<n<1$	$1<n<2$	$n=1$
Total cost $=C=b_1V^n$; $V=K$	Increases monotonically.	Increases exponentially.	Increases linearly.
Marginal cost $=\frac{dC}{dV}=nb_1V^{n-1}$	Positive.	Positive.	Positive equal to b_1.
Slope of marginal cost $=\frac{d^2C}{dV^2}=n(n-1)b_1V^{n-2}$	Negative asymptotically approaches zero.	Positive increases at a decreasing rate.	Zero.

run marginal costs are also independent of capacity.[5] With Equations 3b, 3c, and 3d, however, short-run marginal costs are related to capacity. Equations 3b and 3c specify marginal costs as linear functions of capacity and differ only with respect to the constant term. Equation 3c is forced through the origin. Equation 3d expresses both short-run fixed and marginal costs as quadratic functions of capacity and thus permits, at least, a preliminary exploration of curvilinear relationships.

If b_2 is positive in Equation 3a, as is to be expected, the long-run cost function is defined by setting capacity equal to volume, or:

$$C=a+(b_1+b_2)V$$

For Equations 3b and 3c, short-run marginal costs for any plant with capacity K are given by the partial derivative:

$$\frac{\partial C}{\partial V}=b_2+b_3K$$

we expect b_3 to be negative, reflecting the tendency for larger plants to have lower marginal costs. With these linear formulations, as noted earlier, projections for plants with very large capacity would indicate *negative* short-run marginal costs.

Long-run cost functions for Equations 3b and 3c are defined by setting capacity equal to volume, providing volume is restricted to the range where:

$$\frac{\partial C}{\partial V}=b_1+b_3V>0$$

For the fitted equations, this derivative is negative for annual volumes greater than 235,000 tons for Equation 3b and 245,000 tons for Equation 3c—values substantially smaller than the largest capacity reported for sample plants.

Model 4.—This differs from the previous models in that it has been fitted graphically as an "envelope" function rather than by statistical regression techniques. In brief, plant capacity and plant volume are taken as the base dimensions of a 3-dimensional figure, with annual costs measured vertically above this base. The desired cost function is a surface fitted as an envelope from below to the scatter of individual plant cost points. For convenience in presentation, the resulting surface was represented approximately by an algebraic equation. Finally, a multiple correlation coefficient was calculated as a convenient summary description of the "fit" by comparing deviations between actual and estimated costs with the variance in actual costs. No attempt was made to improve the fit by adjusting the surface, though some effort in this direction would normally be justified and could be expected to yield higher correlation coefficients. Notice that the envelope relationship attempts to define relatively efficient operations with actual costs lower than those achieved by most plants; as a consequence, the correlation coefficient should be somewhat lower than those obtained for average relationships by conventional regression techniques.

The particular approach used to develop this function was (1) to stratify the sample plants on the basis of capacity, (2) to prepare cost-volume scatter diagrams for each strata, and (3) to plot straight-line cost-volume relations for each strata. Each of these straight lines was fitted at or near the bottom of the scatter diagram, and each may be considered an estimate of the short-run total cost function for plants of indicated capacity when designed and operated with reasonable efficiency. The resulting intercepts (fixed costs) and slopes (marginal costs) from the several strata relationships were then plotted against capacity to be "faired" into a smooth surface and finally expressed algebraically in the form:

$$4a: C=b_1K^n+\frac{b_2}{K^m+b_3}V$$

Selection of this particular form was guided entirely by the slope and intercept values from the graphic traces and not by any a priori considerations. The long-run cost function derived from this equation, however, is especially interesting. With relatively small values for V, total cost is minimized by using plants with excess capacity since the increase in fixed costs is more than offset by the reduction in variable costs.[6]

[5] Note that this is equivalent to the Model 2 form with $n=1.0$; an Equation 2e can be obtained from 3a as follows:

$$C=a+(b_1+b_2)V+b_2(K-V)$$

The correlation coefficient for this converted Equation 2e will be the same as for Equation 3a, of course, and can be compared with Equation 1e of the original Phillips form.

[6] An examination of the two terms of the derivative of total cost with respect to K

$$\frac{\partial C}{\partial K}=nb_1K^{n-1}+\frac{-b_2mK^{m-1}}{(K^m+b_3)^2}V$$

indicates that when

(Footnote 6 continued on p. 83.)

Note also that the variable cost term is a reciprocal function of K, thus avoiding the negative marginal costs that make some of the foregoing models questionable in higher volume and capacity ranges.

The Empirical Results

The specific results obtained by applying these several models to the Phillips feed mill data are summarized in table 1. In spite of major differences in form and in the magnitude of computed parameters, all equations "account for" a substantial part of the variance in annual costs for this sample of feed mills—all correlation coefficients are higher than 0.90. Moreover, all V regression coefficients appear to be highly significant (1 percent level) while most of the K and $(K-V)$ coefficients appear to be significant (5 percent level). In general, only the VK, the K^2, and the VK^2 regression coefficients are of doubtful statistical significance. If we followed a rule of omitting from the analysis any variable with a t ratio less than the critical value for a level of significance of 5 percent, for example, Equations 3b and 3c would be reduced to Equation 3a. Equation 3d would be altered sequentially by first dropping K^2, then VK, and finally VK^2, reverting also to Equation 3a.

It is not at all clear, however, that such a rule should be followed. The primary objective in such studies is to estimate the parameters of a cost function, not to test these particular statistical hypotheses. There is a strong a priori reason for expecting that costs will be influenced by capacity (K) and excess capacity ($K-V$), for example, and this is a compelling basis for retaining the $(K-V)$ terms in Equations 1b and 1c even though these regression coefficients are of doubtful statistical significance. Stated in another way, these computed coefficients are the best estimates that the analyses yield for the true values and far better than assuming that the true value is zero.

In a similar sense, if there are good reasons to expect that volume and capacity have a joint effect on costs, then even small values for KV regression coefficients should be retained. The size of the sample may not be large enough to clearly detect differences which, though small in magnitude, may be extremely important in dictating the shape of short- and long-run cost relationships.

In spite of high correlation coefficients and generally acceptable tests of significance for most regression coefficients, many of the equations give results that must be rejected in some ranges. All equations from Models 1 and 2 with n-values of less than 1.0 involve decreasing marginal costs, as indicated earlier, and so may be suspect on logical grounds; these equations also indicate that total costs reach maximum values and then decline, although these points occur well beyond the ranges of actual volumes and capacities. In addition, the negative intercepts for all Model 2 equations must yield unacceptable estimates for low capacity and volume ranges. The cost functions given by Equations 3b and 3c reach maximum values at 235,000 and 245,000 tons, respectively, and so are clearly unacceptable for high volume and capacity estimates. Moreover, Equation 3b has a negative intercept and so must be rejected at least for very small volume and capacity situations. Finally, Equation 3d eventually reaches a maximum although at a figure well beyond actual volume and capacity ranges. The long-run average cost curve based on this equation also has a peculiar form, declining to 35,000 tons, rising to a relative maximum at 185,000 tons, and then declining.

In spite of such logical limitations, there is an implication of almost equal statistical acceptability for the above equations because of the uniformly high coefficients of correlation. The several models, however, yield widely differing estimates of the short- and long-run cost functions. This is illustrated for the long-run functions in figure 1—it would be difficult to devise a more heterogeneous set of relationships, either with respect to the indicated levels of average costs or the rates at which average costs change with increases in scale. Short-run curves are no more consistent, as suggested in figure 2 by average cost relationships for plants with annual capacities equal to 150,000 tons.

(Footnote 6 continued from p. 82.)

$$nb_1K^{n-1} < \frac{b_2mK^{m-1}}{(K^m+b_3)^2}V$$

total variable costs are declining at a more rapid rate than fixed costs are rising as K is increased. Should the above conditions hold at $K=V$, the indication is that V could be produced at lower total cost in a plant with $K>V$. Whether or not these conditions obtain depends upon the value and sign of the parameters of the total cost equation.

Table 1.—*Alternative cost equations derived from identical annual data on total costs, plant volume, and plant capacity, 29 midwestern feed mills*[1]

Model	Total cost equation [2]	Coefficient of correlation
1a	$C=1607.58V^{0.5}+0.73287(K-V)$ (6.84)oo (2.25)o	0.9211
1b	$C=208.98V^{0.7}+0.36436(K-V)$ (12.55)oo (1.73)	0.9695
1c	$C=70.042V^{0.8}+0.30140(K-V)$ (16.80)oo (1.86)	0.9820
1d	$C=22.702V^{0.9}+0.30001(K-V)$ (20.53)oo (2.25)o	0.9876
1e	$C=7.178V^{1.0}+0.34208(K-V)$ (20.79)oo (2.63)o	0.9879
1f	$C=2.229V^{1.1}+0.41178(K-V)$ (18.10)oo (2.83)o	0.9843
2a	$C=-124122+2279.41V^{0.5}+0.6177(K-V)$ (9.98)oo (2.53)o	0.9571
2b	$C=-55423+231.57V^{0.7}+0.4205(K-V)$ (15.13)oo (2.39)o	0.9791
2c	$C=-29863+73.114V^{0.8}+0.3612(K-V)$ (18.09)oo (2.40)o	0.9850
2d	$C=-8281+22.905V^{0.9}+0.3241(K-V)$ (20.26)oo (2.38)o	0.9878
3a	$C=10018+6.8193V+0.3051K$ (15.05)oo (2.27)o	0.9883
3b	$C=-2567+7.1346V+0.4638K-0.00000197VK$ (14.58)oo (2.76)o (1.51)	0.9893
3c	$C=7.1080V+0.4458K-0.00000182VK$ (15.16)oo (3.22)o (1.75)	0.9892
3d	$C=5799+6.5445V+0.2578K+0.00000083K^2$ (6.22)oo (0.43) (0.32) $+0.00000365VK-0.000000000012VK^2$ (0.61) (0.73)	0.9897
4a	$C=0.004122K^{1.4}+109.3V/(K^{0.27}-6.01)$	0.967

[1] Basic data for all models and the results for 1a, 1b, 1c, and 1d were made available by Professor Richard Phillips' Iowa State University, Ames, Iowa.
[2] In all equations, C represents total mill costs in dollars per year, V represents annual mill volume in tons, and K represents computed annual mill capacity in tons. Figures in parentheses are *t* ratios: o indicates significance at 5 percent level, while oo indicates significance at the 1 percent level.

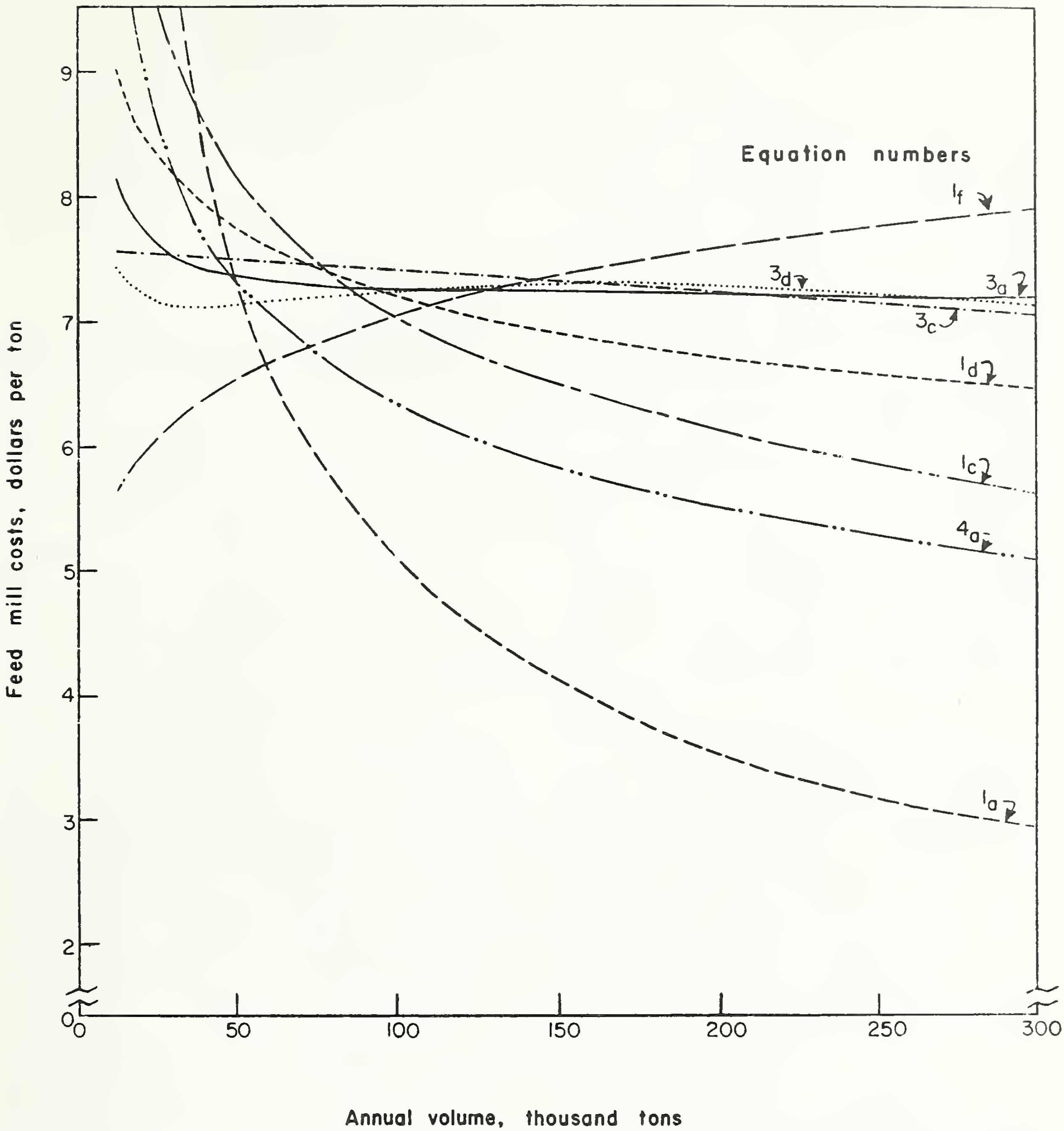

Figure 1. Estimates of economies of scale in feed milling, as derived from eight alternative cost models.

One may well wonder how such completely different results could be obtained from one set of basic data and still yield correlation coefficients and, for the most part, *t*-ratios which suggest high degrees of reliability. In part, this situation can be explained by the fact that changes in equation form were accompanied by compensating changes in the regression coefficients of the independent variables. For example, in the fitted equations for Model 1, there is a systematic inverse relationship between the exponent and multiplier of the volume variable. The changes in the estimated slopes of the regression surface which accompany changes in equation form apparently

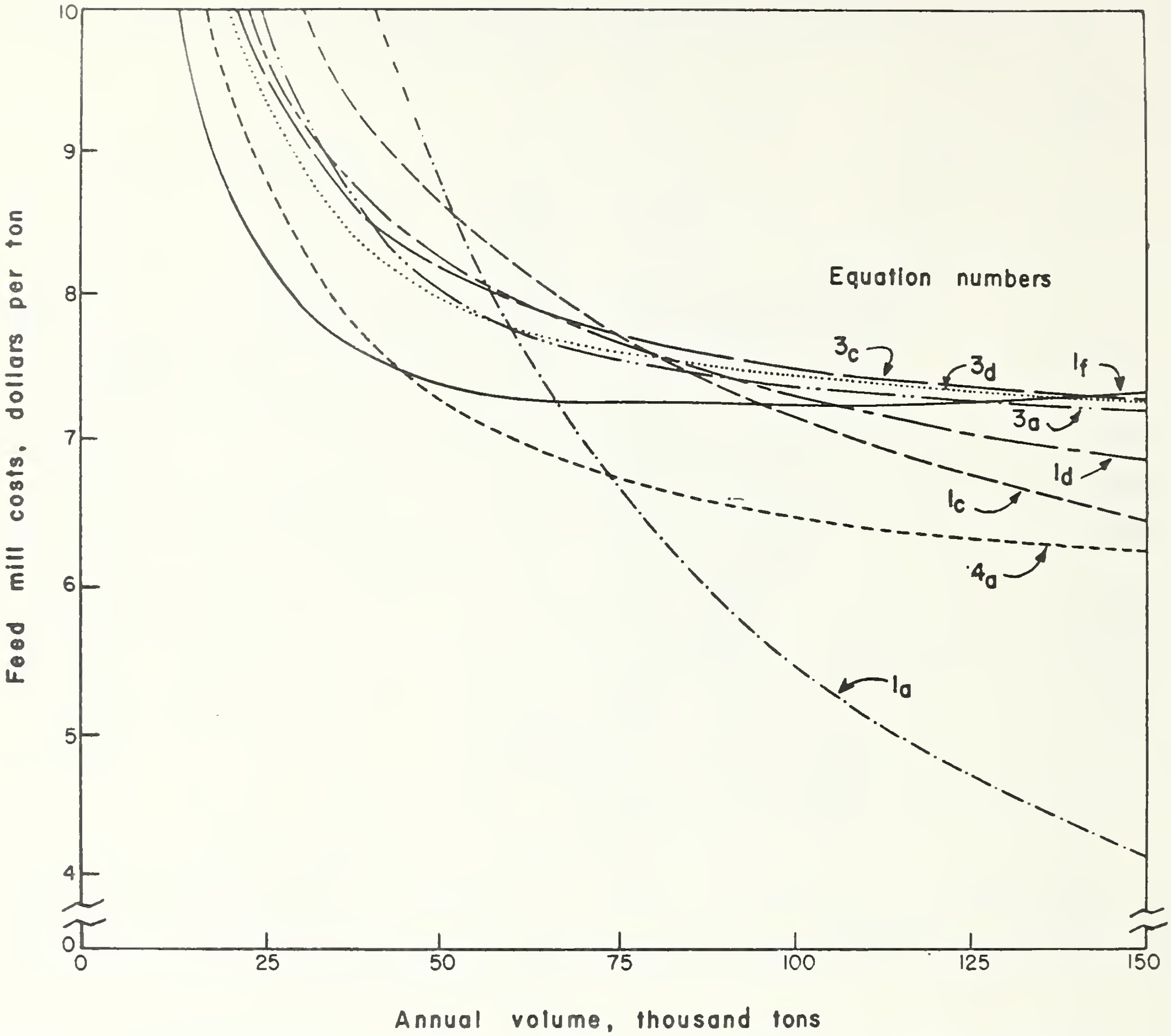

Figure 2. Estimates of short-run average costs for feed mills with annual capacities of 150,000 tons, as derived from eight alternative cost models.

take place in such a way that each of the alternative models fit the observed cost-volume points quite well. However, when these alternative slopes are projected to the long-run situation, the alternative models yield quite different results. The need for caution in projecting the results of any regression analysis is well recognized. However, use of cross-section data to estimate long-run costs will almost invariably involve some form of projection as firms are normally observed at some intermediate point on their short-run cost curves.

An inherent problem in projecting the results of any regression analysis is the lack of certainty that the true slopes of the regression surface have been detected. This lack of certainty prevails even when high multiple correlation coefficients are obtained if the independent variables are highly correlated with one another. In this analysis, the correlation between K and V was 0.856. Intercorrelation not only affects the reliability of the regression coefficients of a given equation but also permits regression surfaces with widely differing slopes in some directions to exhibit uniformly high multiple correlation coefficients.

Conclusions

We have presented the results obtained from the application of several alternative formulations or models of cost relationships. We stress that these formulations have not been selected at random—we are not attempting to demonstrate that a haphazard selection of type equations will yield a haphazard set of regressions. We submit that each general model used has substantial a priori backing; we note that, depending on the specific values obtained for the several regression coefficients, most of the forms used have produced plausible results at least over considerable ranges in volumes and capacities. In short, any one of these formulations might well have been selected by a researcher in attempting to derive quantitative cost functions from cross-section data drawn from a sample of operating plants. While obvious peculiarities in the results for some of the equations would have dictated their rejection, as noted earlier, most would have gratified the research worker by yielding highly "respectable" measures of correlation and of reliability. The individual results certainly seem to justify the assumption that each is a reasonably accurate and dependable description of the true cost functions.

Yet, the wide variety of cost relationships resulting from these trials throws an entirely different light on this matter. Our general conclusion must be that the analysis of such cross-section data may result in high correlations and apparently significant regression coefficients, without providing the basis for confidence in the results as even rough approximations of the basic cost relations involved. It is well recognized that the correlation coefficient is not an adequate guide in selecting among alternative regression forms, and our results emphasize that high and fairly uniform coefficients, plus regression coefficients which for the most part appear to be statistically significant, may be associated with entirely different estimates of the underlying cost functions.

To be specific with respect to this study of feed-mill costs, we are at a loss when faced with the problem of selecting among the several alternative formulations—although we would reject some and limit the range of applicability of others on logical grounds as noted earlier. We do not know whether long-run average costs levels are relatively high or low or if they are characterized by minor declines as scale is increased or by pronounced economies of scale extending over wide ranges in capacity. In a similar way, we find it impossible to forecast the effects of volume on costs for a plant of particular capacity. We would find it difficult or impossible to advise plant owners and managers as to the probable cost consequences of building larger or smaller plants or of combining the volumes for two or three plants in a single operation. Faced by this great diversity of empirical findings, we may well wonder if cost functions derived from cross-section data are fact or fantasy. While these conclusions stem specifically from the analysis of data from a small sample of feed mills, we know that essentially similar situations characterize many other types of marketing and processing plants. We find it difficult to believe, moreover, that the analysis of farm management cross-section data is devoid of such pitfalls.[7]

These somewhat doleful findings do not mean that studies of underlying industry economies of scale and short-run average cost curves based on cross-section data are without value, but they do emphasize that this approach should be used with care and caution. Perhaps, the following generalizations are justified:

1. The usual statistical tests of reliability and of correlation are of very limited usefulness in judging the significance of results as estimates of underlying relationships. The researcher must place primary dependence on a priori reasoning in selecting type equations and even then must be prepared to find that the empirical results are obviously unacceptable in certain ranges; by the same token but, unfortunately, less obvious, the derived relationships are suspect in all ranges as an indication of the underlying structure which determines plant costs.

---

[7] This view is supported by the work of Hildebrand, John R., "Some Difficulties With Empirical Results From Whole-Farm Cobb-Douglas-Type Production Functions," *Journal of Farm Economics*, vol. XLII, November, 1960, pp. 897–904. This work examines the stability of estimated marginal productivities obtained by fitting alternative models to a single set of farm management cross-section data and fittings of the same model to data obtained in three successive years. Hildebrand concludes that ". . . it appears that one can hit on or select a particular model or application of a model to 'support' nearly any recommendation concerning resource use." p. 901.

2. Increasing the size of the sample may to some extent reduce the difficulties encountered in these trials, especially if plants in the sample cover wide ranges in both volume and capacity. It must be recognized, however, that the major problems stem from intercorrelation of the independent variables—here volume and capacity—and that this intercorrelation is not a function of sample size. If our sample plants cover wide ranges in volume for every level of capacity, say from 10 to 100 percent of available capacity, there will be a significant intercorrelation between volume and capacity variables, and this intercorrelation will increase as the sample covers wider and wider ranges in capacity.

3. The intercorrelation between volume and capacity permits compensating shifts in the regression coefficients for these variables; the coefficients are unstable and subject to fairly wide changes in response to chance differences in the plants included in the sample. In essence, these changes represent shifts in the estimates of the relative magnitudes of fixed and variable costs. As a consequence, cross-section data that separates fixed and variable cost components should permit greatly improved cost analyses although the relative levels of fixed and variable costs change markedly with differences in equipment and method.

4. With large samples, the data may be stratified and each stratum analyzed separately along lines similar to those employed for Equation 4a above. If we stratify by capacity, the observations within each stratum are more or less homogeneous with respect to plant size and each observation represents approximately the situation for a plant of this size when operated at the specified volume. Analysis of the strata data, then, should yield good approximations to the short-run cost functions which in turn are traces on the total cost surface.[8]

[8] Studies of cotton ginning costs by W. E. Paulson of the Texas Agricultural Experiment Station are good examples of the possibilities of deriving short-run or plant cost curves from sample strata homogeneous with respect to capacity and type of equipment.

5. As a corollary to (4), we join F. V. Waugh[9] in urging the advantages of graphic analysis or a combination of graphic and more formal methods. Plotting the observations for strata gives the researcher a "feeling" for his data, while visual inspection can be most helpful in selecting a specific equation form within any general a priori model. Moreover, this approach facilitates the use of envelope or near-envelope functions rather than average regressions—a real advantage in many studies.

6. None of the above comments refer to the basic data themselves other than with respect to such components as fixed and variable costs. Since the cost estimates result from accounting records, it is clearly desirable to have all data based on standardized and well-understood accounting systems. Estimates of fixed cost components should be based, ideally, on some standard such as new replacement values; failing this, approximate data on plant and equipment age might permit the inclusion of this factor directly in the analysis. Measures of capacity are especially useful in any attempt to derive both short- and long-run cost relationships, and direct observations of the capacities based on major equipment items should have been better, if available, than the estimates based on past performance used by Phillips. Because of the great importance of seasonal factors, capacity measurements in terms of rates (output per hours, and so on) will usually be most useful. Finally, information on total hours of plant operation may be a strategic addition to the data on plant volume. But these additions to basic information take us further and further away from usual cross-section data and into the area covered by the following paper in this series.

[9] F. V. Waugh, *Graphic Analysis in Agricultural Economics*, U.S. Department of Agriculture Handbook No. 128. Washington, 1957, p. 1.